Data Movement in Distributed Environments: Challenges and Solutions

Raj Kettimuthu

Argonne National Laboratory and

The University of Chicago

Today's Science Environments

- Large-scale collaborative science is becoming increasingly common

Fusion community's International ITER project

- Distributed community of users to access and analyze large amounts of data

the globus alliance
www.globus.org

Large Hadron Collider

1800 Physicists, 150 Institutes, 32 Countries

100 PB of data by 20XX; 50,000 CPUs?

the globus alliance
www.globus.org

Time consuming tasks in Science

- Run experiments
- Collect data
- Manage data
- Move data
- Acquire computers
- Analyze data
- Run simulations
- Compare experiment with simulation
- Search the literature
- Communicate with colleagues
- Publish papers
- Find, configure, install relevant software
- Find, access, analyze relevant data
- Order supplies
- Write proposals
- Write reports
- …

Time consuming tasks in Science

- Run experiments
- Collect data
- Manage data
- Move data
- Acquire computers
- Analyze data
- Run simulations
- Compare experiment with simulation
- Search the literature

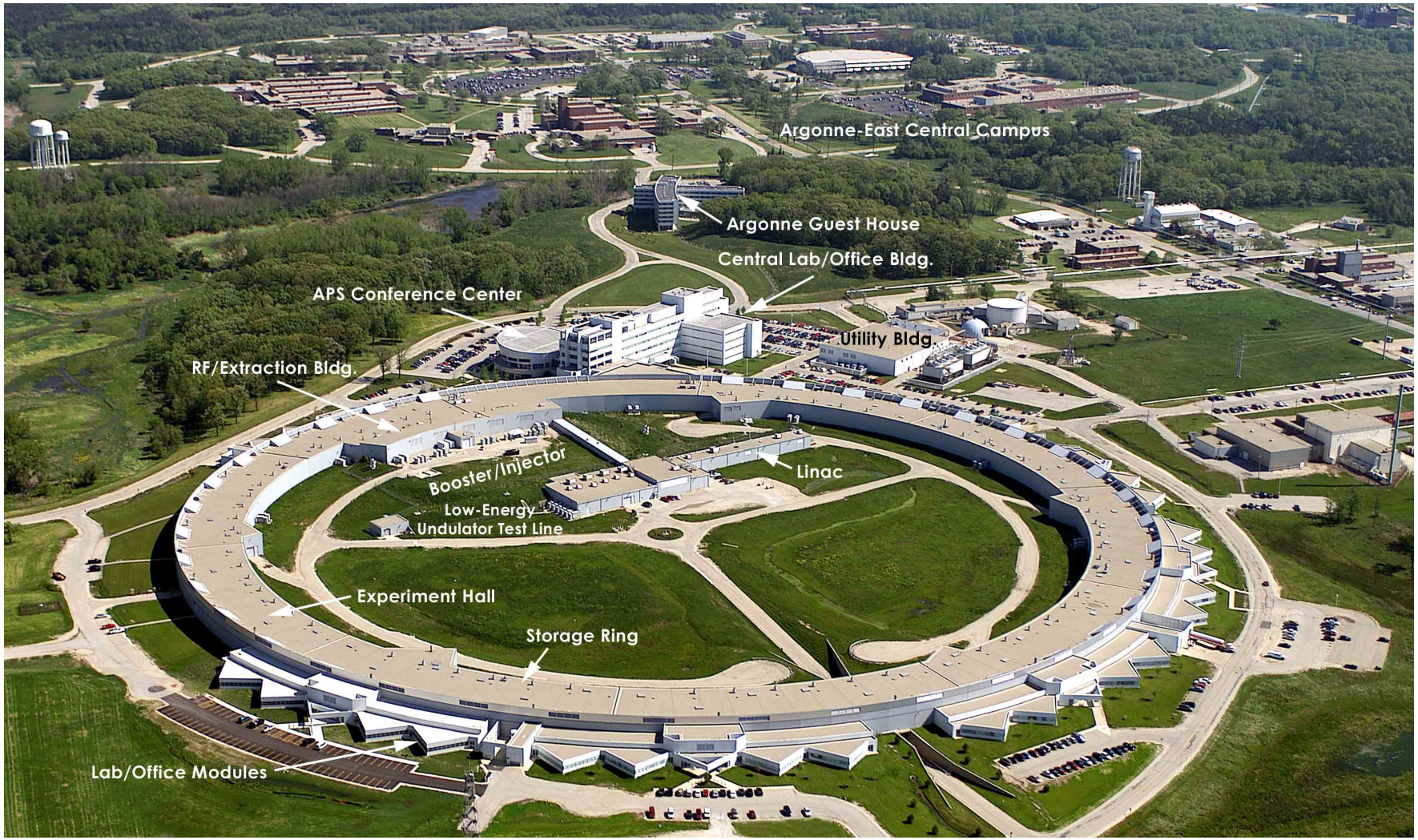
- Communicate with colleagues
- Publish papers
- Find, configure, install relevant software
- Find, access, analyze relevant data
- Order supplies
- Write proposals
- Write reports
- …

Data Movement

- Deceptively simple
- Stick it in email
- Too large
- 100,000 files totaling 10 Terabytes
- Move from a federal laboratory where they were generated to my home institution
- Can be very difficult

Data Movement Challenges

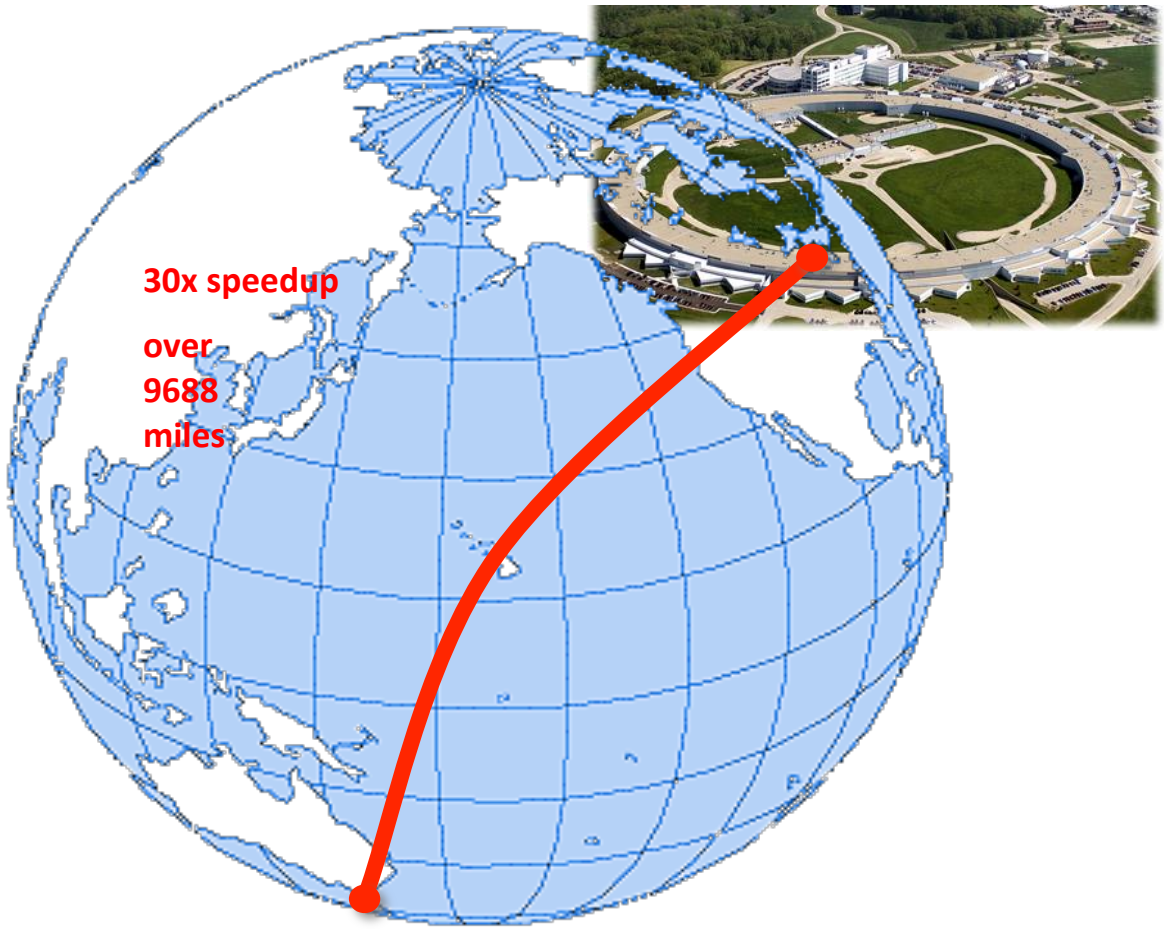
Advanced Photon Source



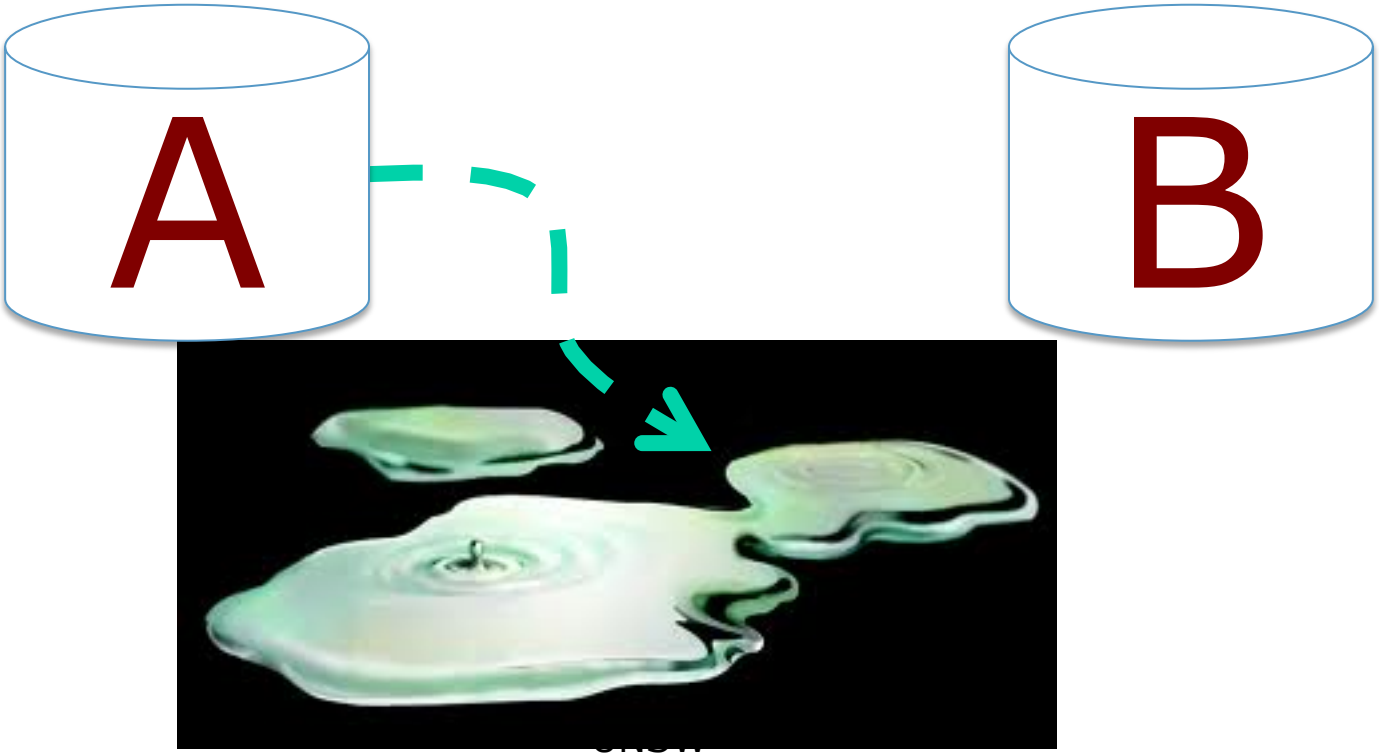
Data Movement Challenges

the globus alliance
www.globus.org

Data Movement Example



the globus alliance
www.globus.org



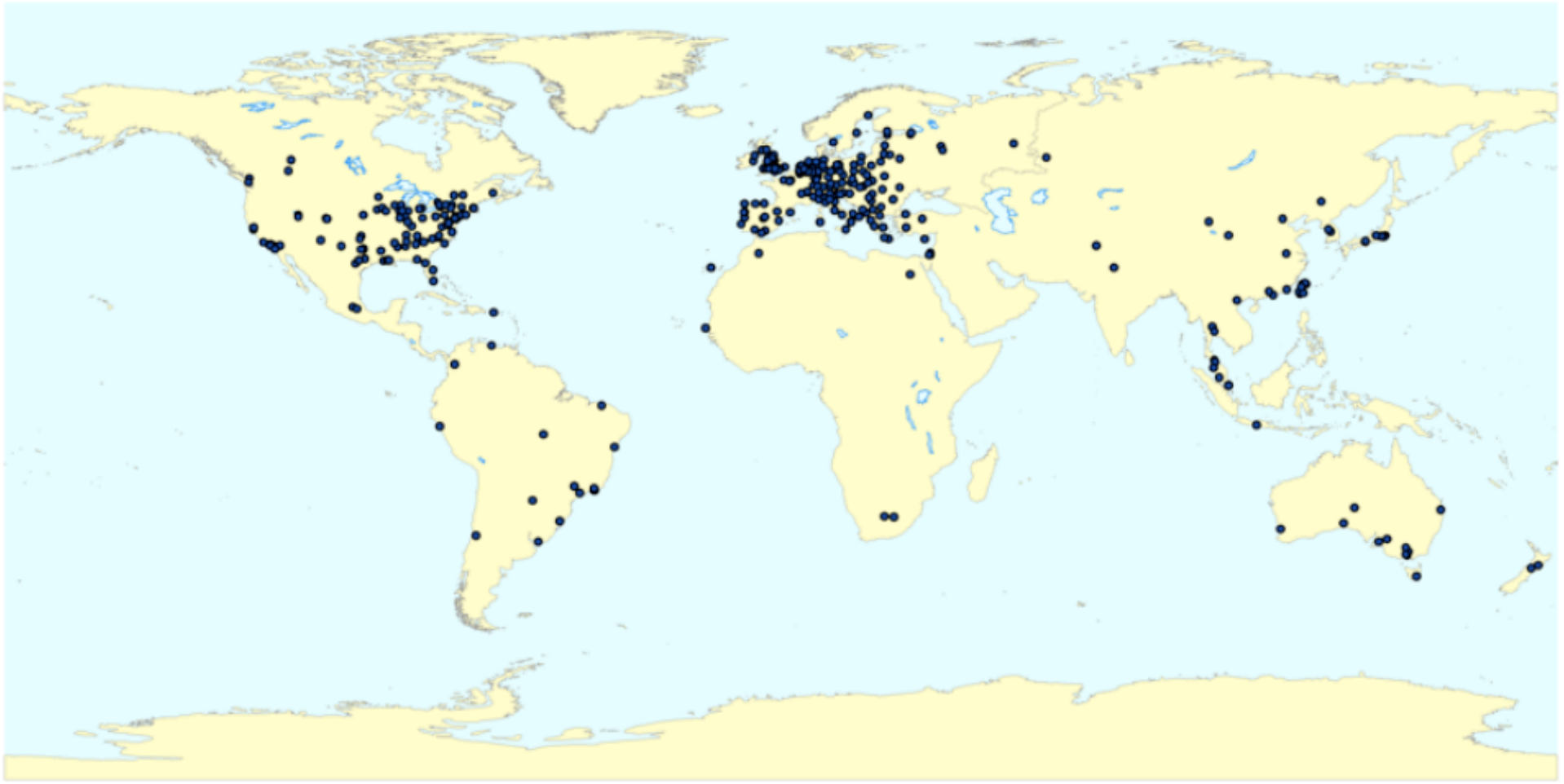
the globus alliance
www.globus.org

GridFTP

- High-performance, reliable data transfer protocol optimized for high-bandwidth wide-area networks
- Globus GridFTP
  - Parallel TCP streams, optimal TCP buffer
  - Non TCP protocol such as UDT
  - Cluster-to-cluster data movement
  - SSH, GSI
  - Restartable transfers

the globus alliance
www.globus.org

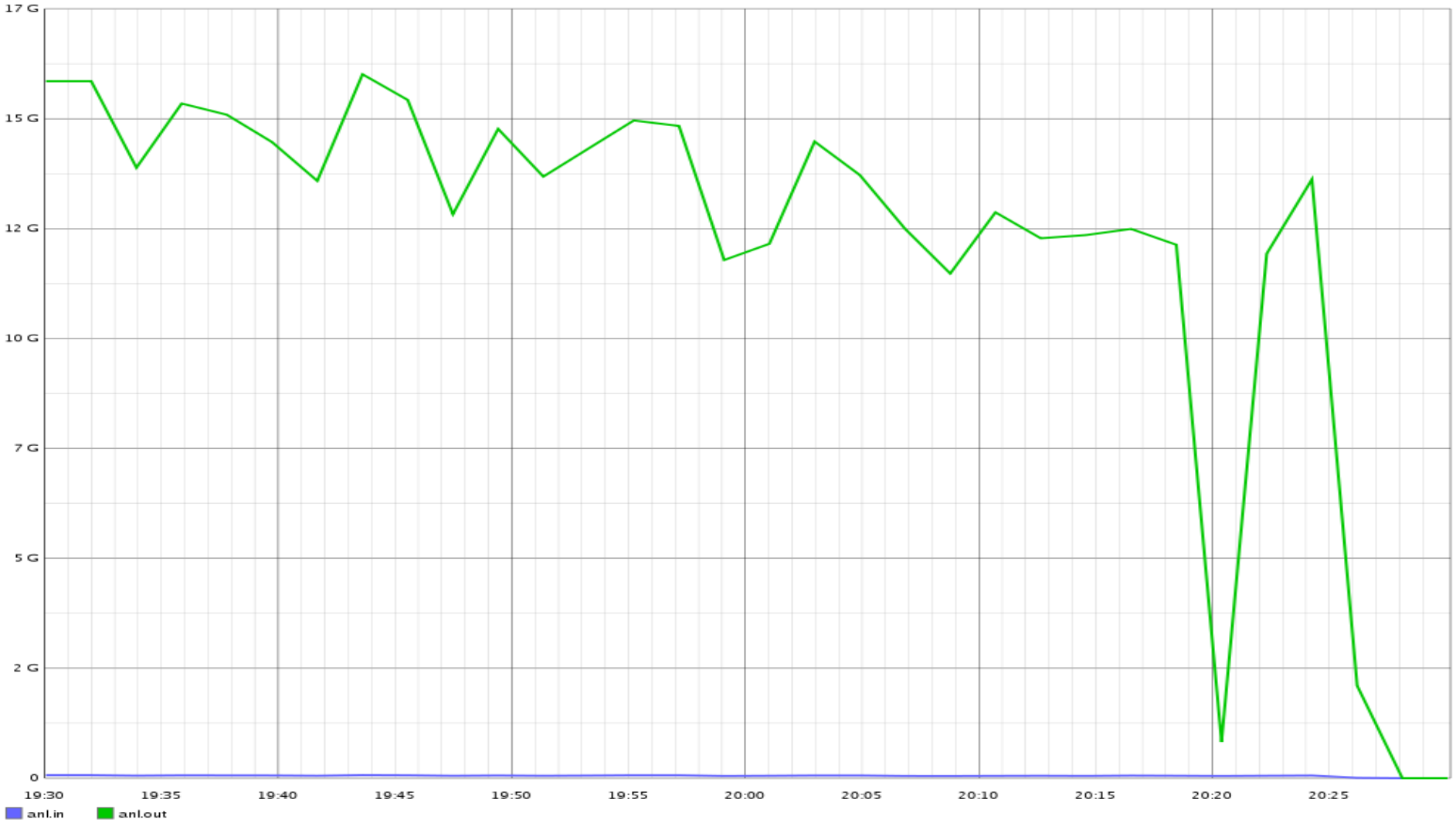
GridFTP Servers Around the World



Created by Tim Pinkawa (Northern Illinois University) using MaxMind's GeoIP technology (http://www.maxmind.com/app/ip-locate).

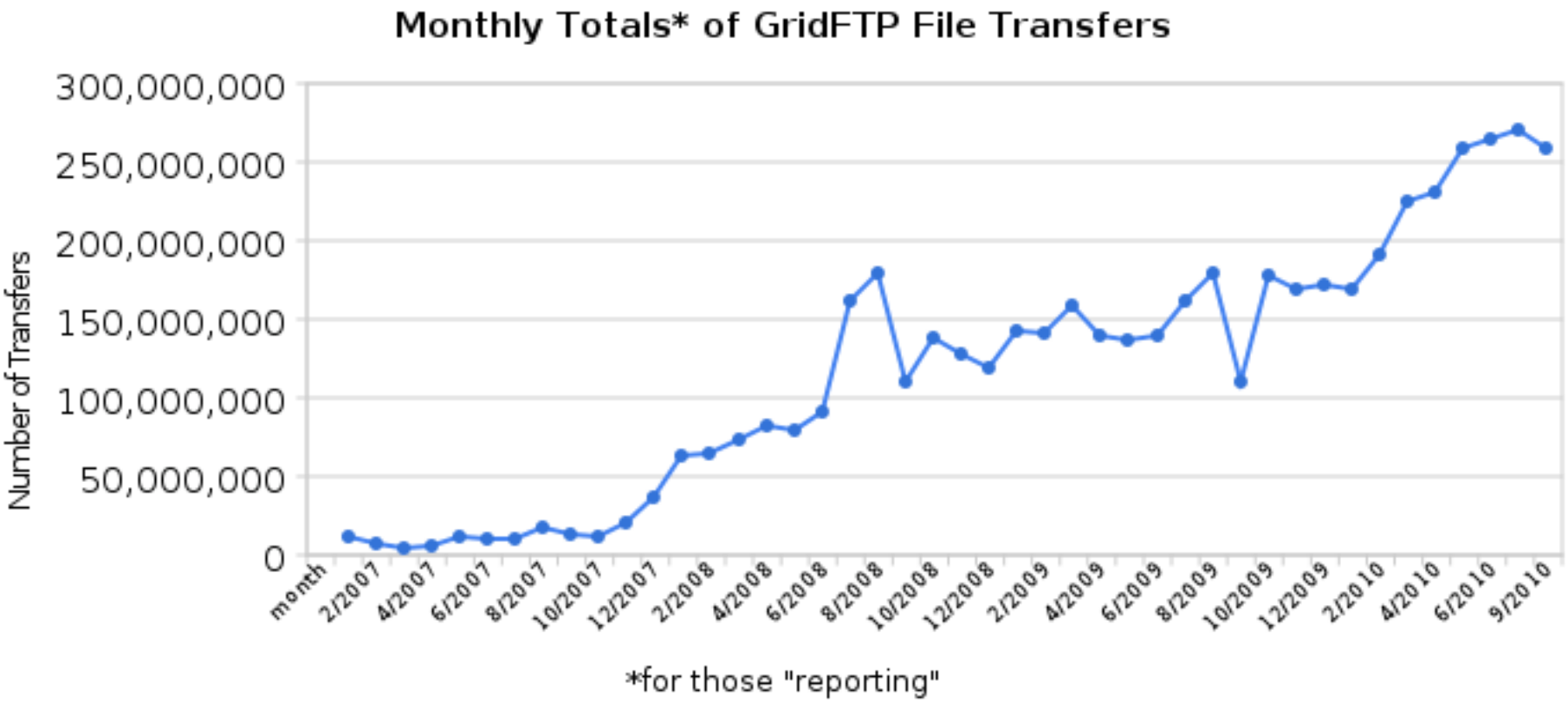
the globus alliance
www.globus.org

Performance



the globus alliance
www.globus.org

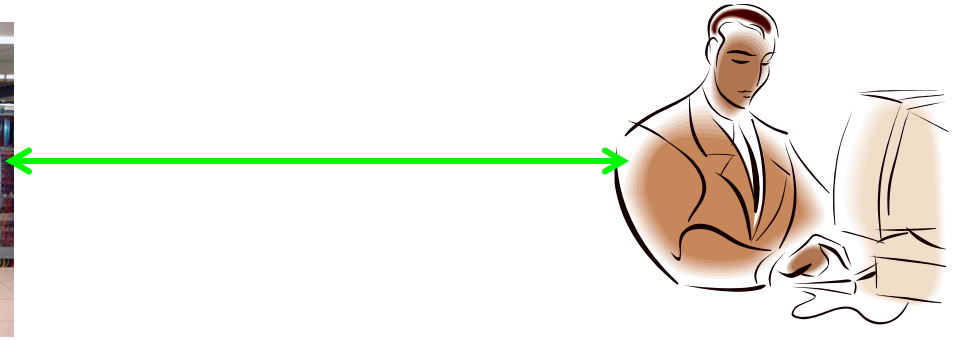
GridFTP Usage



the globus alliance
www.globus.org

GridFTP – user challenges

- Fault recovery
  - Data movement is not a fun activity for users
  - Minimize time spent
  - Reduce user intervention as much as possible
- Installation and configuration

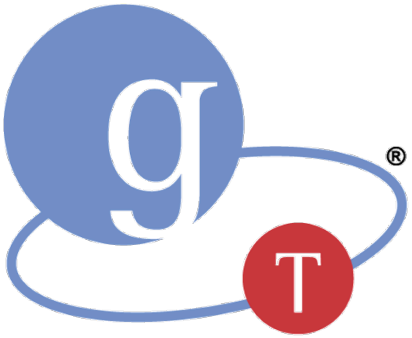


the globus alliance
www.globus.org

Globus Online

- *Globus Toolkit*

Build the Grid



Components for building custom grid solutions

globustoolkit.org

- *Globus Online*

Use the Grid

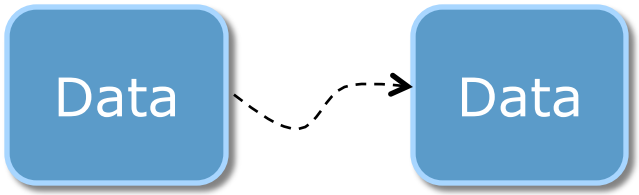
Cloud-hosted file transfer service

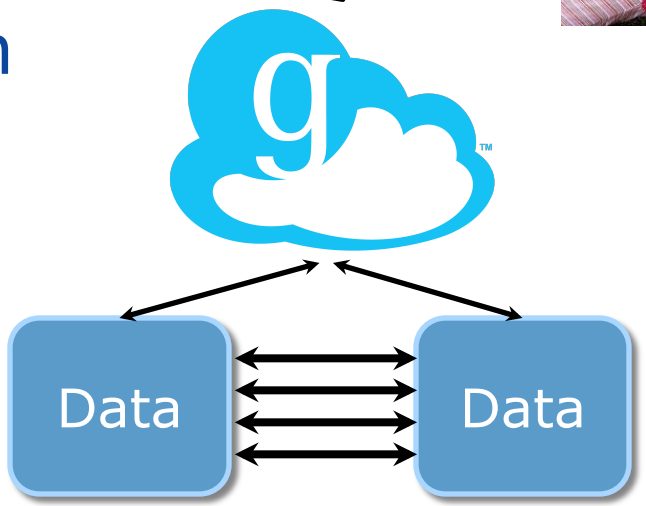
globusonline.org

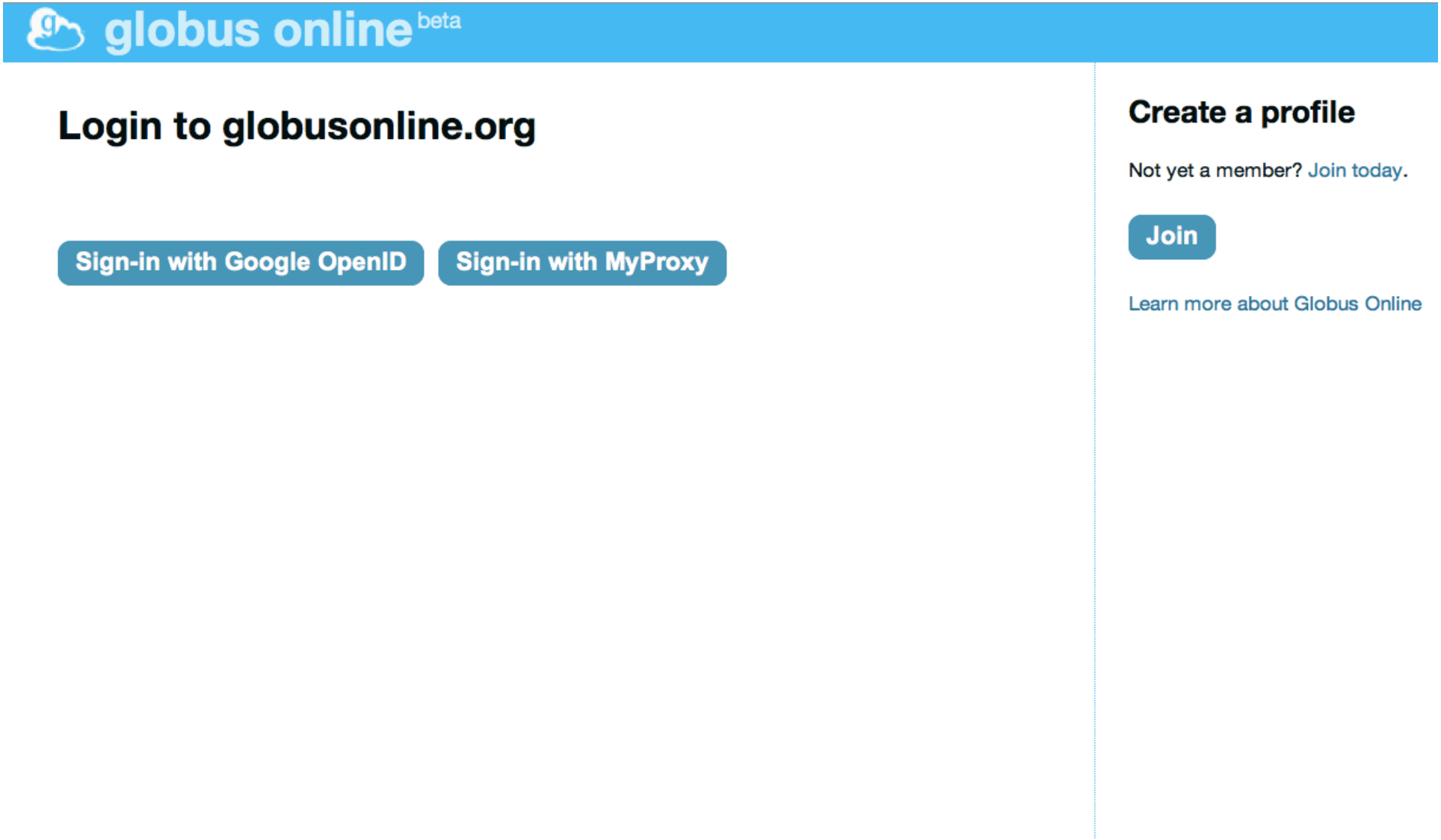
SaaS (Gartner)

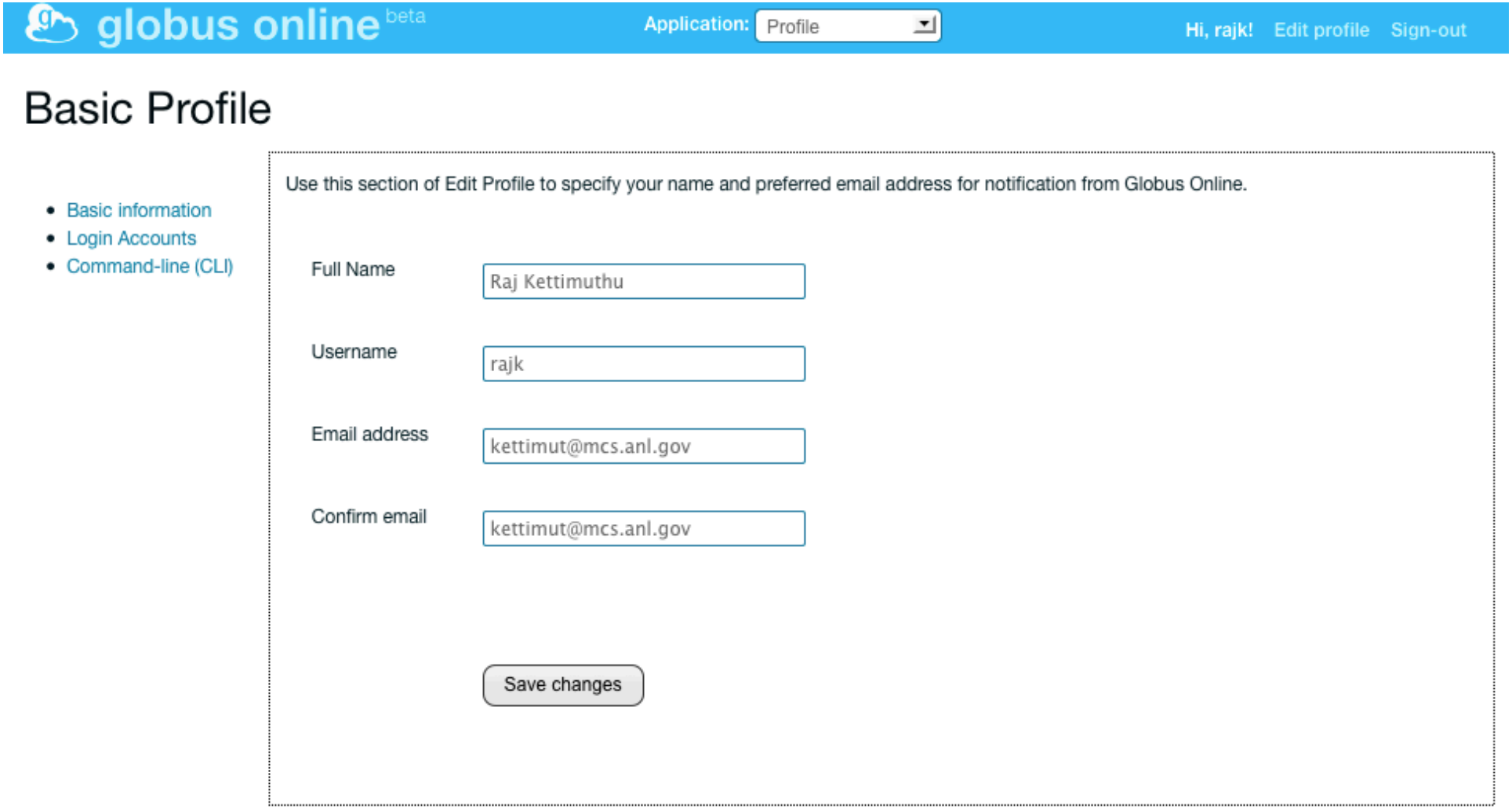
- The application is owned, delivered, and managed remotely by one or more providers
- The application is based on a single code base that is consumed in a one-to-many model by all contracted customers at any time
- The application behind the service is properly web architected—not an existing application web enabled [D. Terrar]

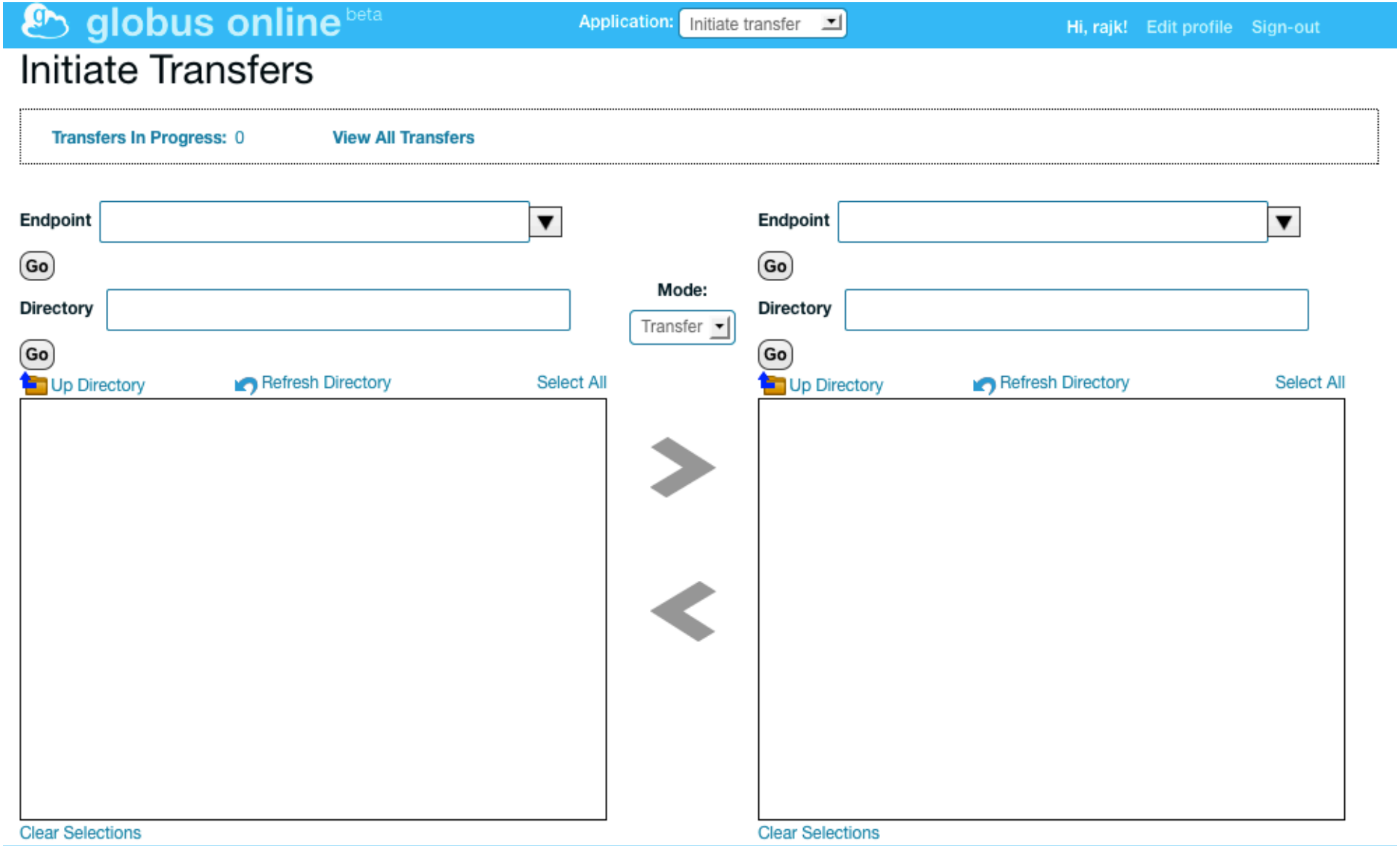
- Easy "fire and forget" file transfers
- Automatic fault recovery
- High performance
- Simplify use of multiple security domains
- No client software installation
- New features automatically available
- Consolidated support and troubleshooting

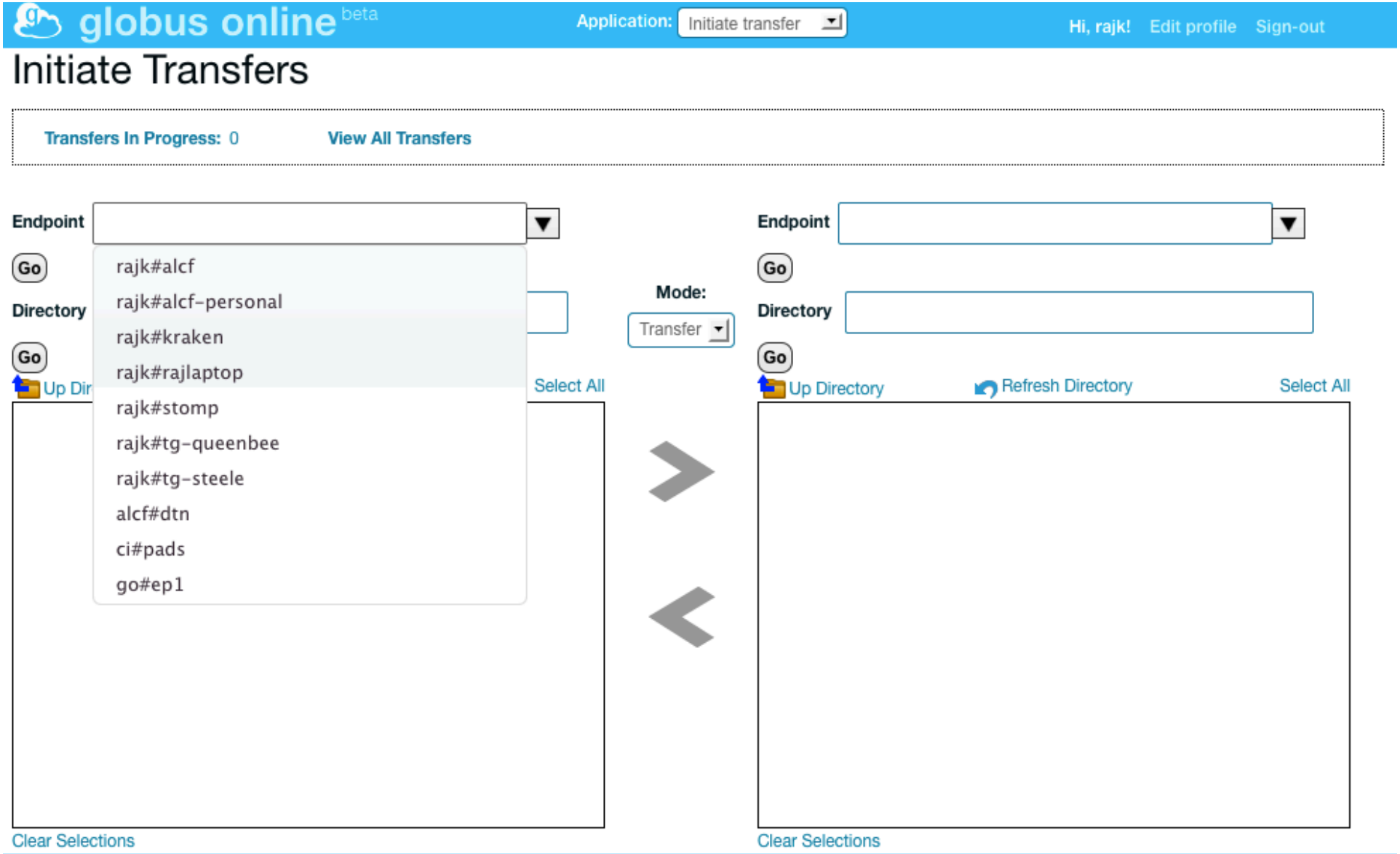


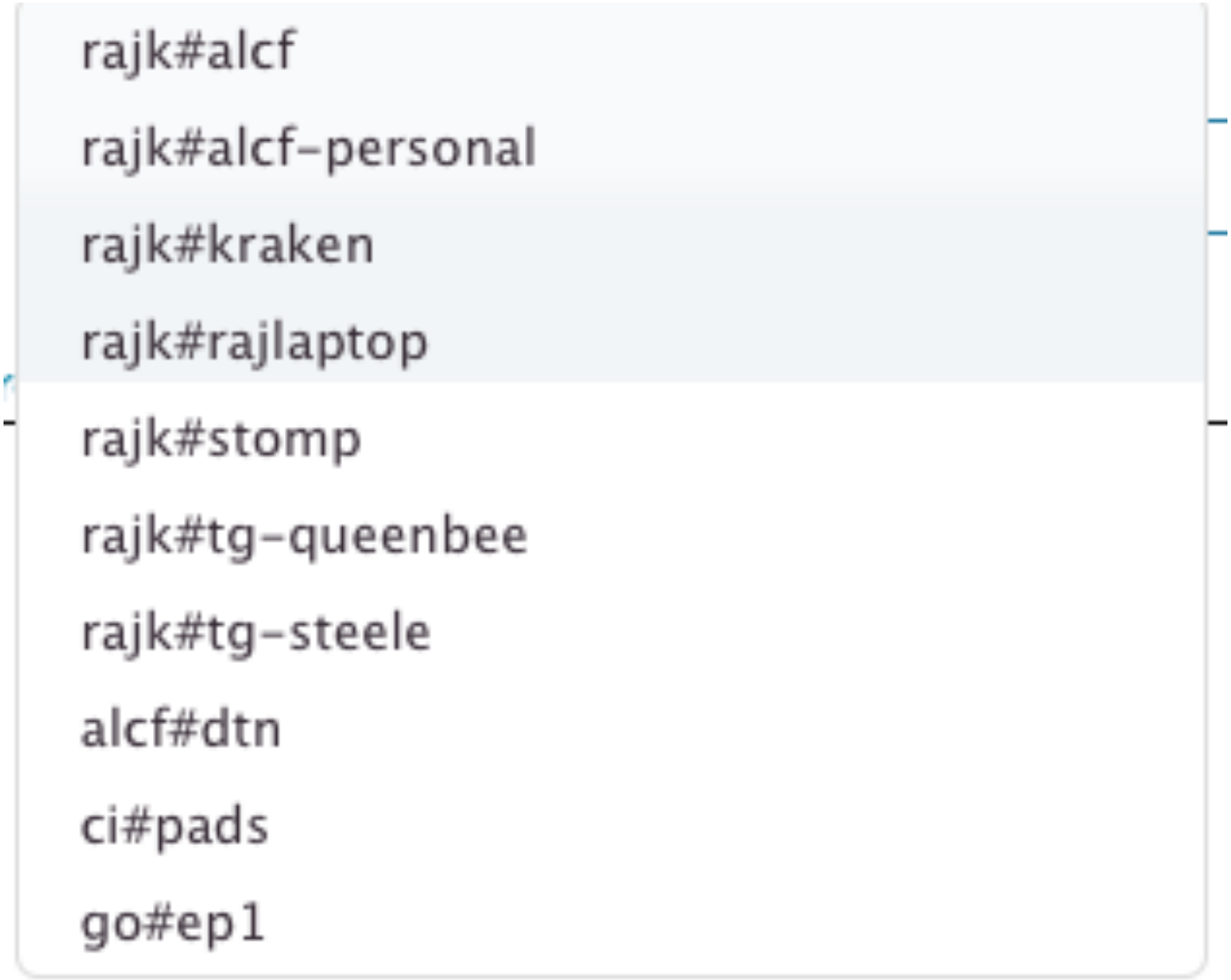



globus online beta
Login to globusonline.org
Sign-in with Google OpenID
Sign-in with MyProxy
Create a profile
Not yet a member? Join today.
Join
Learn more about Globus Online


globus online beta
Application: Profile
Hi, rajk! Edit profile Sign-out
Basic Profile
Basic information
Login Accounts
Command-line (CLI)
Use this section of Edit Profile to specify your name and preferred email address for notification from Globus Online.
Full Name
Raj Kettimuthu
Username
rajk
Email address
kettimut@mcs.anl.gov
Confirm email
kettimut@mcs.anl.gov
Save changes
Globus Online
Overview
Sponsors
Contact Us
Privacy Policy
Terms of Use
Support
Documentation
Video Tutorials
Quickstart
Using the CLI
Beyond Basics
Community
News and Events
Globus Toolkit
Computation Institute
University of Chicago
Argonne


globus online beta
Application: Initiate transfer
Hi, rajk! Edit profile Sign-out
Initiate Transfers
Transfers In Progress: 0
View All Transfers
Endpoint
Go
Directory
Go
Up Directory
Refresh Directory
Select All
Clear Selections
Mode:
Transfer
Endpoint
Go
Directory
Go
Up Directory
Refresh Directory
Select All
Clear Selections
Globus Online
Overview
Support
Documentation
Community
News and Events


globus online beta
Application: Initiate transfer
Hi, rajk! Edit profile Sign-out
Initiate Transfers
Transfers In Progress: 0
View All Transfers
Endpoint
Go
Directory
Go
Up Dir
Select All
rajk#alcf
rajk#alcf-personal
rajk#kraken
rajk#rajlaptop
rajk#stomp
rajk#tg-queenbee
rajk#tg-steele
alcf#dtn
ci#pads
go#ep1
Mode:
Transfer
Endpoint
Go
Directory
Go
Up Directory
Refresh Directory
Select All
Clear Selections
Clear Selections
Globus Online
Overview
Sponsors
Support
Documentation
Video Tutorials
Community
News and Events
Globus Toolkit


rajk#alcf
rajk#alcf-personal
rajk#kraken
rajk#rajlaptop
rajk#stomp
rajk#tg-queenbee
rajk#tg-steele
alcf#dtn
ci#pads
go#ep1

globus online beta Application: Monitor transfers Hi, rajk! Edit profile Sign-out

Monitor Transfers

Remove Data Filter

Cancel ◀◀ ◀ 2 of 2 ▶ 10

	Status	ID	Task Progress	Username	Completion Time	Request Time
☐	!	b0460...	0 / 1 / 1	rajk	12/01/2010 10:55 PM	11/30/2010 10:54 PM
☐	!	ca555...	0 / 1 / 1	rajk	12/01/2010 07:48 PM	11/30/2010 07:47 PM
☐	✓	54282...	1 / 1	rajk	11/18/2010 05:36 PM	11/18/2010 05:36 PM
☐	✓	95a4f...	1 / 1	rajk	11/17/2010 10:11 PM	11/17/2010 10:11 PM
☐	✓	72382...	1 / 1	rajk	11/16/2010 04:54 AM	11/16/2010 04:53 AM
☐	!	27b4b...	0 / 1 / 1	rajk	10/12/2010 05:39 PM	10/12/2010 05:35 PM
☐	!	016b6...	0 / 1 / 1	rajk	10/12/2010 05:29 PM	10/12/2010 05:27 PM
☐	!	9e166...	0 / 1 / 1	rajk	10/01/2010 06:02 PM	10/01/2010 05:59 PM
☐	!	c878c...	31256 / 93 / 31349	rajk	10/03/2010 11:18 PM	10/01/2010 04:34 PM
☐	!	c121f...	0 / 1 / 1	rajk	09/29/2010 10:21 PM	09/29/2010 10:13 PM
☐	✓	56914...	1 / 1	rajk	08/02/2010 07:59 PM	08/02/2010 07:58 PM
☐	✓	eb3c1...	20 / 20	rajk	07/29/2010 04:56 PM	07/29/2010 04:54 PM
☐	✓	ba769...	25 / 25	rajk	07/21/2010 07:09 PM	07/21/2010 06:56 PM
☐	!	8df71...	0 / 40 / 40	rajk	07/21/2010 06:28 PM	07/21/2010 04:18 PM

Cancel ◀◀ ◀ 2 of 2 ▶ 10

CLI 2.0

ssh -t <user>@cli.globusonline.org <command> <options> <params>

Globus Online user name

Globus Online command

Use as needed to hide password text

gsissh <user>@cli.globusonline.org <command> <options> <params>

the globus alliance
www.globus.org

Command Endpoints

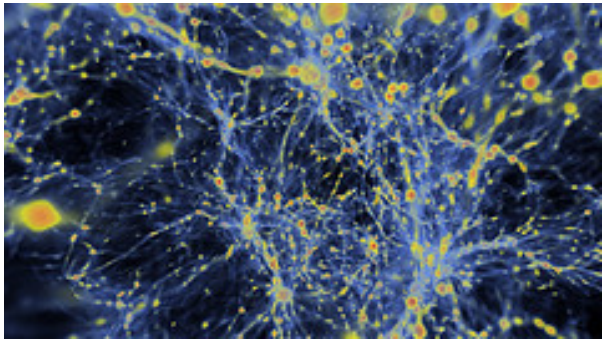

```
ssh lcc@cli.globusonline.org scp go#ep1:/share/godata/file1.txt
go#ep2:~/myfile.txt
Contacting 'myproxy.tutorial.globusonline.org'...
Activating 'ep2'
Activating 'ep1'
Task ID: 19029d64-ecec-11df-aa30-1231350018b1
[XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX] 1/1 0.00
mbps
ssh lcc@cli.globusonline.org ls go#ep2/~/
myfile.txt
```

CLI commands

- cancel
- details
- endpoint-activate
- endpoint-add
- endpoint-deactivate
- endpoint-list
- endpoint-modify
- endpoint-remove
- endpoint-rename
- events
- ls
- profile
- scp
- status
- transfer
- versions
- wait

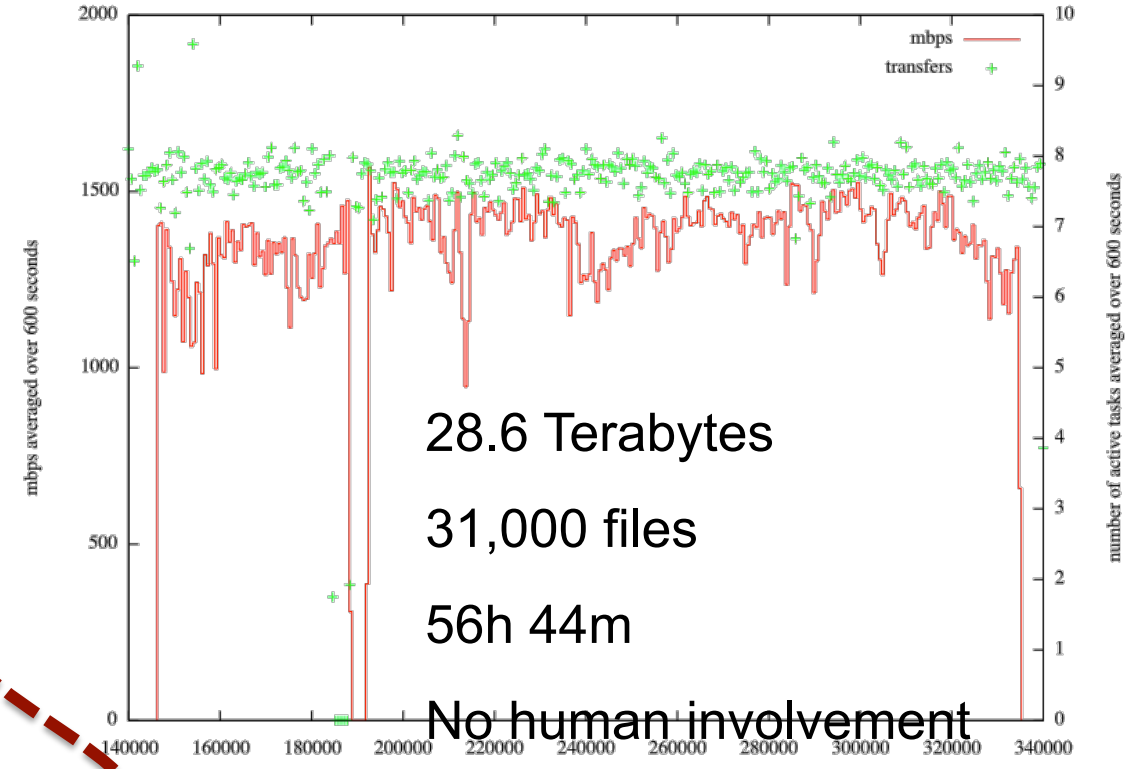
the globus alliance
www.globus.org

Globus Online in Action



Astrophysics simulation data generated in Tennessee, moved to Illinois for visualization

(Enzo, UCSD; Futures Lab, Argonne)



28.6 Terabytes

31,000 files

56h 44m

No human involvement

the globus alliance
www.globus.org

Globus Online

the globus alliance
www.globus.org

REST API

- https://transfer.api.globusonline.org

the globus alliance
www.globus.org

Coming Soon

- Lightweight transfer agent
  - For firewalls, sites without GridFTP installed
- Higher-level data management capabilities
  - Group management
  - Data publication, replication, etc.
  - Workflow
- Additional protocol support
  - HTTP, SRM, …
- Condor integration (version 7.6.0)
  - Stage in and stage out

the globus alliance
www.globus.org

Summary

- Looked at the challenges in data movement for distributed science
- Globus online as one of the solutions
  - Hosted service
  - Easy of use – simple and familiar interfaces
  - Reliability
  - Autotuning for performance
- Looking for users and new collaborators

Acknowledgments

Numerous people have contributed to the Globus Online work, including:

Bryce Allen, Joshua Boverhof, John Bresnahan, Lisa Childers, Paul Dave', Fred Dech, Ian Foster, Dan Gunter, Gopi Kandaswany, Nick Karonis, Raj Kettimuthu, Jack Kordas, Lee Liming, Mike Link, Stu Martin, JP Navarro, Karl Pickett, Mei Hui Su, Steve Tuecke, Vas Vasiliadis

Many thanks to our funders: DOE, NSF, and the University of Chicago

More Information at
http://www.gridftp.org
http://www.globus.org/service/

Questions